DOCUMENT RESUME

ED 482 662 UD 035 962

AUTHOR	Peterson, Paul E.; Howell, William G.
TITLE	Latest Results from the New York City Voucher Experiment.
REPORT NO	PEPG-03-14
PUB DATE	2003-11-00
NOTE	62p.; Paper presented at the Annual Meeting of the Association of Public Policy and Management (Washington, DC, November 2003). Produced by Harvard University, Program on Educational Policy and Governance.
AVAILABLE FROM	For full text: http://www.ksg.harvard.edu/pepg/ index.htm.
PUB TYPE	Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE	EDRS Price MF01/PC03 Plus Postage.
DESCRIPTORS	Academic Achievement; *Black Students; *Educational Vouchers; Elementary Education; Minority Group Children; Private School Aid; *Private Schools; School Choice; Scores; *Test Results; Tuition; *Urban Schools
IDENTIFIERS	New York (New York)

ABSTRACT

The voucher intervention sponsored by New York City's School Choice Scholarships Foundation (SCSF) has been a valuable source of information about such programs. It began in 1997, when K-4 public school students were invited to apply to SCSF for a voucher to help defray costs of private school tuition. Over 1,200 of 20,000 applicants were offered vouchers via a lottery. Since vouchers were awarded by lot, the SCSF program was evaluated as a randomized field trial. Researchers collected baseline test scores and other data prior to the lottery, administered the lottery, then collected followup data 1, 2, and 3 years later. A group families who did not win the lottery acted as the controls. For non-African Americans, and for students as a whole, private schools did not have any discernable impact on test scores. For African Americans, substantial improvements were observed in all 3 years. Although some researchers express strong doubts that African American benefitted much from the New York City voucher intervention, this study suggests that evidence supports the original contention that African Americans, and only African Americans, posted significant and positive test score gains associated with attending a private school that in year 3 ranged from one-quarter to two-fifths of a standard deviation, depending on the model estimated. (Contains 38 references.) (SM)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 482 662

Latest Results from the New York City Voucher Experiment

by

Paul E. Peterson and William G. Howell
Harvard University
PEPG/03-14

Paper prepared for presentation before the
Association of Public Policy and Management
Washington DC November 2003

UD 035 962

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

P. Peterson

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1 2

BEST COPY AVAILABLE

School vouchers are perhaps the most controversial policy reform in education today.[1] Fortunately, in the last decade a wealth of new information has become available about the educational experiences of students in small, targeted voucher programs. The voucher intervention sponsored by the School Choice Scholarships Foundation (SCSF) in New York City has been a valuable source of information about such programs (Howell, Peterson, with Wolf and Campbell, 2003).[2]

This intervention began in the spring of 1997, when those students in grades K–4 who were attending a public school and who were eligible for participation in the free-lunch program were invited to apply to SCSF for a school voucher that would help defray the cost of private-school tuition. More than 20,000 students expressed an interest in the program. Lotteries were held in May, and that fall students began using vouchers to attend private schools. Over 1,200 students were offered vouchers, which were worth up to $1,400 per annum and were initially guaranteed for three years. During the program's first year, 74 percent of families offered vouchers actually used them to send their children to private schools; after two and three years, 62 and 53 percent of the treatment group continued to attend private schools, respectively.[3]

Evaluation Procedures

Since vouchers were awarded by lot, the SCSF program could be evaluated as a randomized field trial. To facilitate the evaluation, the research team collected baseline test scores and other data prior to the lottery, administered the lottery, and then collected follow-up information one, two, and three years later. During the 1997 eligibility verification sessions attended by voucher applicants,

students in grades 1–4 took the Iowa Test of Basic Skills (ITBS) in reading and mathematics.[4] Scheduled during the months of February, March, and April immediately prior to the voucher lottery, sessions were held in private school classrooms, where schoolteachers and administrators served as proctors under the overall supervision of the evaluation team and program sponsors. While children were being tested, accompanying adults completed surveys of their satisfaction with their children's current public schools, their involvement in their children's education, and their demographic characteristics.[5] Over 5,000 students attended baseline sessions in New York City. Mathematica Policy Research (MPR) then administered the lottery in May and SCSF announced the winners.

To assemble a control group, approximately 960 families were randomly selected from those who did not win the lottery.[6] In the absence of administrative error, those offered vouchers should not differ significantly from members of the control group (those who did not win a voucher). Baseline test-score data, easily the best predictor of test-score outcomes (see below), confirm this expectation for those students for whom such data are available. For those students with baseline test scores, therefore, observed differences between the two groups' downstream test scores can safely be attributed to the programmatic intervention.

In the spring of 1998, the annual collection of follow-up information commenced. Testing and questionnaire procedures were similar to those administered at the baseline sessions. Adults accompanying the children again completed surveys that asked a wide range of questions about the educational experiences of their oldest child within the eligible age range. Students completed tests and short questionnaires in schools different from those they were then attending.

To ensure as high a response rate as possible, SCSF conditioned the renewal of scholarships on participation in the evaluation. Members of the control group and students in the treatment group who

initially declined a voucher were compensated for their expenses and told that they could automatically reenter a new lottery if they participated in follow-up sessions. Overall, 82 percent of students in the treatment and control groups attended the year-one follow-up session, as did 66 percent in year two, and 67 percent in year three.

Private-School Impacts on Test Scores

For non–African Americans, and for students taken as a whole, private schools did not have any discernible impact, positive or negative, on test scores. But for African Americans, substantial differences were observed in all three years. African Americans in private schools who were retested after one, two, and three years scored, on average, 6.1, 4.2, and then fully 8.4 National Percentile Rank (NPR) points higher on the combined reading and math portions of the Iowa Test of Basic Skills than their peers in public schools.[7] These findings, furthermore, are robust to numerous alternative specifications and classification schemes. As summarized in Table 1, 108 of 144 different statistical models yield positive and significant effects using a two-tailed test; another 29 are significant using a one-tailed test; the remaining seven are also positive but fall short of conventional levels of statistical significance.[8]

These findings from New York are consistent with those of prior studies using observational data. Surveying the literature on school sector effects and private school vouchers, Princeton Economist Cecilia Rouse says that "the overall impact of private schools is mixed, [but] it does appear that Catholic schools generate higher test scores for African-Americans."[9] Jeffrey Grogger and Derek Neal, economists from the University of Wisconsin and the University of Chicago, respectively, find little in the way of detectable attainment gains for whites, but conclude that "urban minorities in Catholic schools fare much better than similar students in public schools."[10] They are

also consistent with results from experiments in Washington D.C. and Dayton, Ohio, which found no impacts for white students, but, in the second year, positive impacts for African Americans.[11] In the first secondary analysis of the New York experimental data, Barnard, Hill, and Ruben (2003a, p.299) also found, after one year, "positive effects on math scores for children who applied to the program from . . . schools with average test scores below the citywide median. Among these children, the effects are stronger . . . for African American children."[12] And, in the tables of the later secondary analysis by Alan Krueger and Pei Zhu (2004, hereafter KZ), 30 of 51 of the estimations of the voucher impacts on the overall (composite) test scores of African Americans yield significantly positive findings.[13]

Despite the weight of evidence available from the extant literature and from their own estimations, KZ express strong doubts that African Americans benefited from the New York City voucher intervention.[14] At one point in their essay, they suggest "that the provision of vouchers in New York City probably had no more than a trivial effect on the average test performance of participating Black students." In the end, however, KZ back away from this statement, asserting only that "the safest conclusion is probably that the provision of vouchers did not lower the test scores of African Americans"—or, equivalently, that African American students who used vouchers to attend private schools performed as well or better than their peers in public school.[15]

How do KZ generate findings that justify their conclusion? Three analytical decisions stand out as most consequential: 1) Include students without baseline scores in the analysis, despite the risk of obtaining a biased estimate of the program's effects; 2) Employ an unusual, questionable ethnic classification scheme; and 3) Add 28 additional variables to the statistical models, despite their own admitted warnings against "specification searching," rummaging theoretically barefoot through data in the hopes of finding desired results.

The mere addition of students without baseline scores—the analytic decision that KZ claim to be the "most important" evidence in support of null findings—does not, by itself, provide a basis for their conclusions. Results remain significantly positive for African American students in all three outcome years when these students are added to the study. Nor do results change materially if one takes a second step upon which KZ place great weight, the reclassification of students as African American when either their mother or their father is African American. When these observations are added to the sample, estimated voucher effects for African- American test scores remain significantly positive.[16]

If these methodological innovations do not, by themselves, significantly alter the results, they are nonetheless problematic for reasons discussed below. For these and other reasons, we remain convinced that the evidence supports our original contention that African Americans, and only African Americans, posted significant and positive test score gains associated with attending a private school that in year three ranged from one quarter to two fifths of a standard deviation, depending upon the model estimated.

Issue #1: How Important Are Baseline Test Scores?

In a study of student achievement, of all information to be collected at baseline, the most critical is test scores. As stated in a project proposal prepared before outcome data had been collected, "The math and reading achievement tests completed by students [at baseline] will provide a benchmark against which to compare students' future test scores" (Corporation for the Advancement of Policy Evaluation with Mathematica Policy Research, Inc., 1997).[17] More than any other information collected, baseline test scores have the highest correlations with test score outcomes—0.7, 0.6, and 0.7 for years one, two and three, respectively. None of the correlations logged by demographic variables is even half as large.[18]

Unfortunately, Mathematica Policy Research (MPR), the firm that administered the evaluation, was not able to obtain test-score data for everyone at baseline. Some students in grades 1–4 were sick, others refused to take the test, and some tests were lost in the administrative process.[19] And due to the substantial difficulties of testing students who lacked reading skills, no kindergartners were tested at baseline.[20]

So as to follow the original research plan and use the highest quality data, Howell and Peterson with Wolf and Campbell (2002) examined voucher impacts on students for whom benchmark test score data were available. For African American students with available baseline test scores (the Available Tests at Baseline, or the ATBs), one observes moderately large impacts of attending a private school on the combined math and reading portions of the Iowa Test of Basic Skills.[21] Effects are 6.1, 4.2, and 8.4 percentile points in years one, two and three—all of which are statistically significant (see Table 2, row 1).[22] The estimated impacts of private-school attendance on test scores remains significantly positive when students without baseline test scores (No Available Tests at Baseline or NATBs) are added to the analysis.[23] The magnitude of the estimations, however, attenuates because the test scores of African American NATBs were affected either trivially or negatively by attending a private school. For African American NATBs, impacts are 0.1, -3.5, and -13.3 NPR points in years one, two, and three respectively.

The differences in results for the ATBs and the NATBs are sufficiently striking to raise questions about the credibility of the data for the latter group. Consider the following thought experiment: two randomized experiments are conducted, one for a larger number of cases with baseline test scores, the other for fewer cases without this crucial baseline information. The two studies yield noticeably different results. Which of the two should be given greater weight by policy

analysts? If the experiments were of equal quality in other respects, we doubt any scientist would give greater credence to the study set lacking such crucial baseline information.

The thought experiment is a useful exercise because it underscores the fact that concerns about bias arise whenever key baseline information is missing. For ATBs, we have solid grounds for concluding that estimations are unbiased, simply because we know the treatment and control groups do not differ significantly in their baseline test scores. Only a minuscule, statistically insignificant 0.4 NPR points differentiate the composite baseline scores of African American students in the treatment and control groups.[24] But if there seems to be little danger of bias among ATBs, the same cannot be said for NATBs, which may have initially been—or subsequently became—significantly unbalanced. KZ argue otherwise, saying that "because of random assignment . . . estimates are unbiased." But estimates are unbiased only if the randomization process worked as well for the NATBs as it did for the ATBs—an outcome that KZ assume but cannot show (Peterson and Howell 2003, note 19). In the words of the statisticians who first conducted a secondary analysis of the New York experiment, KZ's "assertion that 'because assignment to treatment status was random . . ., a simple comparison of means between treatments and controls without conditioning on baseline scores provides an unbiased estimate of the average treatment effect' is simply false, because there are missing outcomes" (Barnard et al 2003b, 321).

There are a variety of attributes of the New York experiment that make the KZ claim, if not false, then at least exceedingly problematic. The administration of the New York experiment was quite complicated, as KZ themselves lament. Half the sample was selected by means of a matching propensity score design, half by stratified sampling that took into account the date students took the test, the quality of the public school they came from, and the size of the family applying for a voucher. Because many more students and families came to the testing sessions than were

eventually included in the control group, lotteries proceeded in two steps: lottery winners first were drawn randomly, and then a second sample was drawn randomly from non-winners for inclusion in the experiment.

For ATBs taken as a whole, we know that administrative complications did not generate significant test-score differences at baseline. Unfortunately, no information on this crucial point is available for the NATBs. We do know, however, that along a variety of other dimensions (whether a student came from an under-performing public school, the student's gender, and whether the mother graduated from college), significant differences between NATBs in the treatment and control groups are observed. Whether these imbalances extend to NATB test scores to their baseline scores is impossible to know.

Baseline test score imbalances among NATBs may be especially likely among those students in the experiment who were assigned to treatment and control conditions using the matched propensity design, which relied upon baseline test scores whenever they were available. Among the NATBs, student assignments were made only on the basis of available demographic data; and because these data are weakly correlated with outcome test scores, they make for fragile indicators when constructing adequate treatment and control group (Barnard et al. 2000a, 300).

Beyond the creation of the treatment and control groups, additional administrative errors may have occurred. For one thing, matching student names from one year to the next presented numerous complications. For ATB students, the risk of mismatching was reduced because students put their own names on the baseline test and all subsequent tests they took. But for NATBs, student identification at baseline could be obtained only from parent surveys, which then had to be matched with information the child gave on tests taken in subsequent years. NATB parents, furthermore,

were less likely to complete survey questionnaires than ATB parents. Background information is missing for 38 percent of NATBs, as compared to 29 percent of ATBs.[25]

The seemingly mundane job of matching students actually presented multiple challenges. In a low-income, urban, predominantly single-parent population, children's surnames often do not match that of both their parents; children may take their mother's maiden name, their father's name, the name of a stepfather, or of someone else altogether. Also, students may report one or another nickname on follow-up tests, while parents report the student's formal name. Without documentation completed by students at baseline, ample opportunities arise for mismatching parent survey information at baseline and child self-identification in years one, two, and three—raising further doubts about the reliability of the NATB data.[26]

Finally, attrition from the experiment introduces additional risks of bias, risks that lead Barnard et al. (2003a) to characterize the experiment as "broken."[27] When baseline scores are not available, one simply does not know whether this attrition compromised the baseline test score balance between the two groups. For all these reasons, estimates are best made for students for whom baseline test scores are available.

For the moment, though, let us set aside the possibilities of bias arising due to problems encountered during sample construction, administrative error, or differential attrition. What, exactly, is to be gained from introducing the NATBs to the analysis? KZ suggest two potential benefits: the ability to generalize findings to another grade level (kindergartners) and the efficiency gains associated with estimating models with larger sample sizes. On the former score, the kindergartners appear to be quite different from their older peers, making any such generalization hazardous. African American students in grades 1-4 posted significant and positive test score gains (whether or not one includes the NATBs in the analysis, and whether or not controls for baseline test scores are

included) in all three years. Impacts for kindergartners, meanwhile, were more erratic, bottoming out at -13.9 in year three.[28]

At first glance, however, KZ appear justified when espousing the benefits of enlarging the number of available observations. All else equal, the precision of estimated impacts increases with sample size. The problem, of course, is that all else is not equal. And the efficiency gains associated with increasing the number of observations do not make up for the losses associated with not being able to control for baseline test scores.[29] Among African American ATBs, the standard errors for impacts in years one, two, and three in test score models that do not include baseline test scores are 2.3, 2.4, and 3.3 (see Table 2, row 2).[30] When controls for baseline test scores are added, the standard errors drop noticeably to 1.7, 2.2, and 2.9 for the three years (Table 2, row 1). When expanding the sample to include both ATBs and NATBs and dropping controls for baseline test scores, the standard errors jump up to 2.1, 2.2, and 3.0 (Table 2, row 4) As the English would put it, what is gained on the straightaway is more than lost on the roundabouts.

When including students without baseline scores, KZ (2003) reports only the imprecise model.[31] By contrast, the initial secondary analysis (Barnard et al. 2003a) included baseline test scores, whenever possible, in order to obtain as precise an estimate as possible. In a series of estimations KZ 2004 follows suit and controls for baseline scores, though without estimating the model in a transparent manner that allows for straightforward comparisons with the impacts originally reported. Instead, the hybrid model is estimated only after recoding the ethnic identity of some African Americans and adding numerous other demographic controls and missing-data indicators (on these issues, see below). When one does estimate a simple, transparent hybrid model that just controls for baseline test scores, whenever possible, results are only marginally different from those originally reported (see Table 2, rows 5 and 6).[32] To generate findings that justify their

conclusion that vouchers had insignificant effects on African American students, KZ cannot simply add students without baseline scores to the estimations. Instead, they must make additional methodological moves, the next being the introduction of a flawed ethnic classification scheme.

Issue #2: Who Is African American?

In the New York evaluation, families' ethnic backgrounds were ascertained from information provided in the parent questionnaire.[33] At baseline (and, again, at the year-two and year-three follow-up sessions), accompanying adults were asked to place the student's mother and, separately, the student's father into one of the following ethnic groups: 1) Black/African American (non-Hispanic); 2) White (non-Hispanic); 3) Puerto Rican; 4) Dominican; 5) Other Hispanic (Cuban, Mexican, Chicano, or other Latin American); 6) American Indian or Alaskan Native; 7) Chinese; 8) Other Asian or Pacific Islander (Japanese, Korean, Filipino, Vietnamese, Cambodian, Indian/Pakistani, or other Asian); 9) Other (Write in: _____).

Students of "other" background. In most instances, one can easily infer each student's ethnicity simply based on the ethnicity of the parents, as indicated by the responses to this question on the survey. For some cases, however, judgment is required. Should those classified as "other" be reclassified into one of the listed categories? If so, which category? Much, of course, depends upon whether a parent selected the "other" category intentionally or inadvertently. For example, if respondents checked "other" but then claimed to be "Hispanic," it seems safe to assume that they overlooked the Hispanic category above, making reclassification appropriate. The same applies for anyone who inadvertently checked "other" but listed themselves as "African American" or "black." If, however, the "other" category appears chosen with some clear intention, then the respondent was left in that category.

At baseline, the ethnic background of 78 mothers and 73 fathers was identified as "other." Among those students for whom test score information is available beyond the baseline year, *none* of these parents can be reclassified as African American simply because a clear mistake was made by those completing the survey.[34] Rather, these parents identified themselves, quite intentionally, as "black-Haitian," "Puerto Rican/black," "black-West Indies," "black-Cuban American," and "black/Jamaica." Because none of these parents identified themselves simply as "African American" or "black," the safest classification decision is to preserve their self-identification as "other."[35]

KZ (2003, p. 317, Table 2) nonetheless reclassify parents of those in the "other" category as "Black, non-Hispanic" even when the respondents themselves have rejected that label.[36] But it is misleading—and contrary to the very federal guidelines that Krueger and Zhu use to bolster their case—to classify as "Black, non-Hispanic" people who openly identify themselves as "Hispanic," "Dominican," or "West Indian."

According to the federal guidelines KZ cite, a person is to be defined as "Hispanic" if she is "of Mexican, Puerto Rican, Cuban, Central or South American or other Spanish culture or origin, regardless of race," while "a person is 'black'" if she is from "any of the black racial groups of Africa." The guidelines go on to say that if a "combined format is used to collect racial and ethnic data, the minimum acceptable categories are 'Black, not of Hispanic Origin,' 'Hispanic,' and 'White, not of Hispanic Origin,'" adding further that "any reporting . . . which uses more detail shall be organized in such a way that the additional categories can be aggregated into these basic racial/ethnic categories."[37]

To defend their classification of some Hispanics as "black non-Hispanic," KZ (2004) cite studies that indicate that "society treats individuals with different skin tones differently," a point that

Krueger made more starkly when he identified the dark-skinned Dominican baseball player, Sammy Sosa, as "black" when displaying his picture in his National Press Club presentation of the KZ (2004) paper.[38] But the point to be taken away from this image is not that Sosa is "black" but that ethnicity does not reduce to "skin tones."[39] The "skin tones" of many Hispanic students in New York City are just as dark as those of many African Americans (just as the "skin tones" of many African Americans are as light as those of other ethnic groups, e.g., Pacific Islanders, Pakistanis, or Indians). Nothing in OMB's Statistical Directive 15 says that Hispanics should be classified according to their skin color or any other physical attribute. To the contrary, the Directive says that if "race and ethnicity are collected separately, the number of White and Black persons who are Hispanic must be identifiable, and capable of being reported in that category."

KZ (2002) employed a probit model to estimate the percentage of Dominicans thought to be black, and then used the results of the model to recalculate voucher effects, which were not significant when these estimated black Dominicans were included in the model. Actual results from these models were dropped in KZ 2003 and KZ 2004, but the basic idea of re-classifying Hispanic students as black, non-Hispanic persists (see, for example, KZ 2004). We are unaware of scholarly precedents for this classification system.

Students of Mixed Ethnic Heritage. According to OMB's Statistical Directive 15, persons who are of mixed racial and/or ethnic origins should be placed in the category "which most closely reflects the individual's recognition in his community." The procedure we employed—classifying students by the ethnicity of the mother—is certainly consistent with the guideline, for the simple reason that in the overwhelming percentage of cases the mother is the person with whom the child lives. However, the guidelines might also be interpreted as allowing for the classification of students

according to the ethnicity of the mother and father, taken together, or of the primary parental caretaker.

Eschewing these alternatives, KZ employ a unique classification scheme. They identify students of mixed heritage as African American, as long as either the mother or the father is African American. If a child has a mother who is Hispanic, but a father who is African American, KZ classify the child as "black, non-Hispanic."[40] As a consequence, students cannot be classified as Hispanic (while maintaining mutually exclusive categories) unless neither parent is African American. KZ defend this classification scheme on the grounds that it is "symmetrical." But symmetry is hardly the word for a scheme that classifies Hispanics and African Americans according to different principles.

Howell and Peterson with Campbell and Wolf (2002) classify all students according to a single principle—students consistently were assigned to their mother's ethnic identification, a procedure also used by Barnard et al (2003).[41] Since it is a child's mother who strongly influences the educational outcomes of most low-income, inner-city children, it is the schooling options available to these mothers that matter most.[42] Several items in the parent questionnaire demonstrate the primary role that mothers played in the lives of the students participating in the study. Of the 792 ATB students with African American mothers who were tested in at least one subsequent year, 67 percent lived with their mother only, as compared to just 1 percent who lived only with their father.[43] The mothers of 74 percent of these students were single, divorced, separated, or widowed; in fact, only 20 percent of the children lived in families where the mother was married. Mothers accompanied 84 percent of children to testing sessions; and in 94 percent of the cases, the accompanying adult claimed to be a caretaker of the child. All of these factors point in the same direction—mothers, as an empirical fact, were most responsible for the educational setting in which

the children in this study were raised. Since the educational choices available to the mother are what matter most for the child, students should be classified according to her ethnicity.[44]

With this in mind, we show results in Table 3 from four classification schemes. The first three represent classification schemes that are consistent with federal guidelines.[45] First, as done originally, the students' ethnic background is defined by the mothers'. Second, students are identified as African American, if both parents are. Third, the child's ethnicity is identified by the ethnicity of the parental caretaker (most frequently the mother, but occasionally the father). In all three years, and for all three of these plausible classification schemes, the same results emerge: private-school impacts on the test scores of African Americans, however defined, are positive and significant (see columns 1–3, Table 3).

Nor do the results change materially when students are identified as African American if either their father or their mother is African American.[46] Although inconsistent, this decision, by itself, is not sufficient to reach conclusions different from those originally reported. For all students with and without baseline test scores, statistically significant, positive impacts on African Americans are estimated in all three years (see column 4, Table 3).[47]

Issue #3: Which Covariates Should Be Included in the Analysis?

Using hybrid models that take into account baseline scores whenever possible, we have shown significantly positive impacts of private schooling on the test scores of all participating African American students (defined in various ways). KZ do not report these simple, transparent estimates. Instead, in KZ (2002), hybrid models include 12 other regressors (8 family and student characteristics and 4 missing variable indicators). KZ (2004) adds 16 more (8 characteristics and 8 missing data indicators).[48]

The decision to add all of these covariates obviously forsakes the values of simplicity and parsimony (see, for example, Zellner 1984, p. 31). Unfortunately, it also provides little gain in the precision of the estimates obtained (see below). Equally important, it increases the chances of introducing bias. First, when adding covariates, KZ impute means and include indicator variables to denote cases with missing values. In doing so, KZ must make the highly restrictive assumption that neither the background variables nor missing-value indicators correlate with treatment; for if they do, then the estimated treatment effects may be biased.[49] As Achen (1986, p. 27) points out, when working with less-than-perfect randomized experiments, "controlling for additional variables in a regression may worsen the estimate of the treatment effect, even when the additional variables improve the specification,"[50] a problem KZ themselves admit: "if there is a chance difference in a baseline characteristic between treatments and controls, there could also be an erroneous correlation (due to chance or misspecification) between the baseline characteristic and the outcome variable that would sway the estimated treatment effect if covariates are included."[51]

Given such risks, a good rule of thumb is to avoid adding a covariate unless treatment and control groups are shown to be balanced and significant gains in precision are achieved. As previously shown, inclusion of benchmark test scores passes both of these tests: baseline test scores of treatment and control groups remained balanced from baseline to the year three study; and the inclusion of baseline test scores as covariates substantially improves the precision of estimated treatment effects. [52] The same, however, cannot be said for the 28 additional covariates that KZ introduce to the analysis.[53]

Elsewhere in their essay, KZ themselves express doubts about models that include background controls. As they put it,

> Estimates without baseline covariates are simple and transparent. And unless the specific covariates that are to be controlled are fully described in advance of analyzing the data in a

> project proposal or planning document, there is always the possibility of specification searching.

This argument suggests that only baseline scores, the one variable identified in the project proposal as theoretically relevant, should be included in statistical models that estimate achievement gains. Inasmuch as additional background controls were not introduced from the beginning of the research project, it is problematic to add them now.

The rules set forth by KZ, of course, apply to secondary analyses as well. Whenever possible, researchers should identify in advance the covariates to be included in their statistical models, especially when these covariates can artificially inflate or deflate the estimates. And when lists of covariates change over time—compare KZ 2002 with KZ 2004—questions naturally arise about the possibility of specification searching.

To show how results change when covariates are added, Table 4 reports third-year private-school impacts that control for different numbers of background control variables, for different classifications of African Americans, and for students with and without baseline test scores. Columns 1–4 report estimated impacts for ATBs; columns 5–8 report impacts for ATBs and NATBs together. For those African American students with baseline scores, the results do not change significantly when covariates are added (see columns 1–4). No matter how many additional regressors are successively added to the statistical models, positive and statistically significant impacts emerge.

Inclusion of new covariates changes results only when the NATBs are added to the analysis (see columns 5–8). Even then, estimated impacts for two of the four definitions of ethnicity remain significant on a two-tailed test when the first seven background variables are included; and all estimations remain significant on a one-tailed test when one adds the covariates originally identified by KZ (2002) to be relevant. Only when still further background characteristics are introduced do

the effects of private-school attendance attenuate—though on a one-tailed test, estimates still are significant for every definition of African American except the novel one proposed by KZ. Unfortunately, with the addition of each new background characteristic, one after another, one repeatedly makes the restrictive assumption that all students with missing data are alike with regards to the item in question.

Since the inclusion of additional covariates requires strong assumptions, one should avoid them unless they add materially to the precision of the estimate. In this instance, it is not even a close call. Among the ATBs and NATBs, the inclusion of additional covariates never reduces standard errors by more than a minuscule 0.05 NPR percentile points. Indeed, the addition of these covariates actually causes standard errors to increase in two of the four definitions of African American background. Far from providing a more "powerful" estimate, as KZ have claimed, the addition of all these variables frequently has the opposite effect.[54]

Concluding Observations

The findings reported by Howell and Peterson with Wolf and Campbell (2002) are robust to a wide variety of alternative specifications and classifications. Only in a very few models do the results fall short of significance at conventional levels (see Table 1). Importantly, the few models that do not yield statistically significant results are the most restrictive in that they suffer from at least two of the following difficulties: 1) large numbers of students for whom no baseline data were available were introduced into the analysis; 2) a novel, inconsistent ethnic classification scheme was employed; and 3) the analysts, without *ex ante* theoretical justification and after conducting at least two separate specification searches, added to the model 28 covariates for which much information is missing. In our view, there is no basis for privileging these estimations over the many others that have a superior scientific foundation.

What, then, can be learned of more general significance from this further analysis of the New York voucher experiment? The following come to mind:

1) Randomized experiments yield data that are less threatened by selection bias than most observational studies, but they are usually difficult undertakings in which administrative error is possible and sample attrition likely. To verify an experiment's integrity, baseline data on the key characteristic one is measuring are vital.

2) A randomized field trial is not strengthened by introducing observations that potentially disrupt the balance between treatment and control groups.

3) When classifying students, mutually exclusive categories should be employed and equivalent coding rules that follow standard practice should apply to students of different ethnic backgrounds.

4) In randomized field trials, covariates should only be added when treatment and control groups are shown to be balanced, and significant gains in precision are achieved.

For these reasons, we conclude that the weight of the evidence from the evaluation of the New York voucher intervention lends further support to the finding—found repeatedly in both experimental and observational studies—that poor African American students living in urban environments benefit from private schooling.

ENDNOTES

[1] The many groups and individuals who assisted with the evaluation are acknowledged in Howell, Peterson with Wolf and Campbell (2002). Here we wish to thank as well those who have provided comments on this paper, including Alan Altshuler, Christopher Berry, David E. Campbell, Morris Fiorina, Alan Gerber, Donald Green, Jay Greene, Erik Hanushek, Frederick Hess, Caroline Minter Hoxby, Martin West, and Patrick Wolf. Howell, Peterson, with Wolf and Campbell (2002) also includes findings from voucher experiments in other cities. Also, see Howell, Wolf, Campbell, & Peterson 2002 and Peterson, Howell, Wolf & Campbell 2003.

[2] This study also includes information from voucher experiments elsewhere.

[3] In all three years, a small percentage of the control group also attended private schools.

[4] The assessment used in this study is Form M of the Iowa Test of Basic Skills, Copyright 1996 by The University of Iowa, published by The Riverside Publishing Company, 425 Spring Lake Drive, Itasca, Illinois 60143-2079. All rights reserved. The producer of the ITBS graded the tests.

[5] For a comprehensive analysis of these data, see Howell, Peterson, Wolf, and Campbell (2002).

[6] Exact procedures for the formation of the control group are described in Hill, Rubin, and Thomas, 2002.

[7] Estimates here differ slightly from those originally reported because MPR, after certifying an original set of weights and lottery indicators in Mayer, Peterson, Myers, Tuttle and Howell (2002),

revised them in 2003. In total, 622, 497, and 519 African American students without baseline test scores were included in test score models after years one, two, and three, respectively. Model specifications provided in Table 2.

[8] Given the sheer number of studies of private schools that find positive achievement and attainment effects for African American students (Chubb and Moe 1990; Coleman, Hofer, and Kilgore 1982; Evans and Schwab 1993; Figlio and Stone 1999; Grogger and Neal 2000; Jencks 1985; Neal 1997; Rouse 2000), and given that no major study—to our knowledge—has found that private schools adversely affect the education of African American students, a one-tailed test is not inappropriate.

[9] Rouse 2000, p. 19.

[10] Grogger and Neal, 2000, p. 153. See works cited in note 8 above as well as Sanders 2002. As one reviewer of the Sander's book notes, "When both [experimental and non-experimental] types of studies yield similar conclusions, the results inspire greater confidence (Godwin, 2002, 83).

In Milwaukee, positive impacts of vouchers on student test scores were observed in an experimental study, most clearly after three and four years. Greene, Peterson, and Du, 1998. In this randomized field trial, baseline test scores were available for only 29 percent of the voucher students and 49 percent of the control group—just 83 students after three years and 31 students after four years, making it extremely difficult to detect effects, positive or negative. As a result, the researchers placed greater weight on data from all students (300 in the third year, 112 in the fourth), whether or not baseline information was available (pp. 345-48). All results were positive, though at various levels of significance. Nonetheless wary of the problem missing benchmark scores posed, the authors pointed out that "the conclusions that can be drawn from our study are . . . restricted by

limitations of the data. . . . . The percentage of missing cases is especially large when one introduces controls for . . . pre-experimental test scores. But given the consistency and magnitude of the findings . . . they suggest the desirability of further randomized experiments capable of reaching more precise estimates of efficiency gains through privatization. Randomized experiments are underway in New York City, Dayton, and Washington, D.C. If the evaluations of these randomized experiments minimize the number of missing cases and collect pre-experimental data for all subjects. . . , they could . . . provide more precise estimates of potential efficiency gains" (p. 351).

[11]In Washington, D.C., however, no statistically significant effects for African Americans were observed in year three. For details and additional results, see Howell, Peterson, with Wolf and Campbell (2002); Howell, Wolf, Campbell, & Peterson (2002) and Peterson, Howell, Wolf & Campbell (2003).

[12] Their analysis uses a statistical method that attempts to adjust for missing cases and non-compliance. Barnard et al examine the effects of the intervention on only those students who came from families with but one child participating in the program. Despite the differences in sample composition and methodological approach, their findings resemble those that we have reported. Over 85 percent of the African American students in the Barnard et al (2003a) analysis came from public schools with average test scores below the citywide median. As the authors point out, "the positive effects . . . for children originating from low-applicant schools are primarily attributable to gains among the African-American children (p. 310)." Within the sample Barnard et al. examined, only 58 African American students came from public schools above the median, a small number from which to draw conclusions.

It is not clear why Barnard et al. (2003a) distinguish between students coming from higher and lower performing public schools, since reported differences between the two groups appear not to be statistically significant. By contrast, differences between African American students and other students are statistically significant (Howell and Peterson with Campbell and Wolf 2002).

[13] If not otherwise identified, all references in this paper are to KZ 2004.

[14] KZ's essay focuses on a narrow band of the research reported in Howell, Peterson, with Wolf and Campbell (2002). KZ do not question the results from the parent surveys, which showed that private schools have lower levels of fighting, cheating, property destruction, absenteeism, tardiness, and racial conflict; assign more homework; establish more extensive communications with parents; contain fewer students and smaller classes; and provide fewer resources and more limited facilities. Nor do KZ question certain null findings, namely that the voucher programs did not consistently increase parental involvement with their child's education, that they had little effect on children's self-esteem, and that they did not adversely impact the degree of racial integration in school.

[15] A preference for "safe" estimates implicitly favors the status quo. In Krueger's view, "Policymakers should be risk-averse when it comes to changing public school systems" (as quoted in Neal 2003).

[16] Effects are significant according to a two-tail test, in all years for students with baseline test scores. If students without baseline scores are included in the analysis, results are significant in years one and three according to two-tail test and second year results are significant according to

one-tail test.

[17] This document was prepared roughly five months prior to the beginning of the collection of outcome data.

[18] A few other characteristics—mother's education, entry into grade 4, learning disabled student, gifted student, and Protestant religious affiliation—register significant correlations with test score outcomes in all three outcome years. Their correlations, however, never exceed 0.25.

[19] Twenty-four African American students (or 10.6 percent of the sample) in grade 1, 34 (12.9 percent) in grade 2, 21 (8.9 percent) in grade 3, and 25 (13.6 percent) in grade 4 had missing baseline test scores. All 245 African American kindergartners had missing baseline test scores. According to the original research proposal, MPR, the firm responsible for data collection, was to include in the lottery only those students in grades 1–4 for whom baseline test score information was available. As stated in the proposal, "The second phase of the application process will include completing a questionnaire with items that ask parents . . . to describe the basic demographic characteristics of the families. In addition, MPR will administer a standardized achievement test to students and ask students to complete a short questionnaire . . . . Children will be excluded from the lottery if they do not complete the . . . application process." (Corporation for the Advancement of Policy Evaluation with Mathematic Policy Research, Inc. 1996.) After the lottery was held, MPR reported that administrative procedures were not fully executed according to plan, as some students for whom no baseline test scores were available nonetheless were given a chance to win a voucher. Also, MPR did not make some of the test score data to the propensity score matching team until after their work was completed, causing problems with the construction of the control group (Barnard et

al. 2003a, 301).

[20] Parent surveys and tests were administered to all students in subsequent years; to do otherwise would have drawn distinctions among children and families, inviting suspicion among the participants.

[21] KZ (2004) report results for composite scores as well as for the math and reading portions of the test, separately. Composite scores yield more precise estimations, however; their standard errors are 15 to 20 percent lower. Given these efficiency gains, we report only impacts on composite test scores. Krueger (1999) employed this analytical strategy in his reanalysis of data from the Tennessee class size study, even when precision was less of an issue, as the number of cases available for observation totaled around 10,000 students.

[22] Weighted, 2SLS regressions estimated where treatment status is used as an instrument. As covariates, models for the ATB group include private school status, baseline test scores, and lottery indicators. For the NATB group, covariates only include private school status and lottery indicators.

Estimates of private school impacts compare those students who attended a private school for three years to those students who did not. If students benefited from attending a private school for one or two years and then returned to a public school, this approach will overstate the programmatic impacts. On the other hand, if switching back and forth between public and private schools negatively impacts student achievement, then this model will underestimate the true impact of consistent private-school attendance.

When KZ estimate two-stage models, they assume that private school impacts accrue at a linear rate. (Nothing about the models we estimate imposes an assumption that gains must be linear.) Still, whether one estimates impacts one way or another is not particularly consequential. Our third-year estimated impact for ATB students is 8.4 NPR points; KZ's is 6.4 points. Although KZ stress that they show voucher impacts that are 31 percent less than the size of the impacts originally estimated, this appears a rather forced interpretation of the finding. Both estimates are statistically significant, and neither is significantly different from the other.

KZ (2004) argue that programmatic effects are best understood by examining the impact of being offered a voucher rather than the impact of actually attending a private school. The first impact, known as intent-to-treat (ITT), is estimated by ordinary least squares (OLS); the second by a two-stage model (2SLS), which uses the randomized assignment to treatment and control conditions as an instrument for private school attendance. Almost all of the estimates KZ provide are based on the OLS model.

To ascertain the statistical significance of programmatic effects, it makes no difference which model is estimated. Both yield identical results. If, however, one is interested in the magnitude of an intervention's impact, not just its statistical significance, then the choice of models is critical. The two estimators will yield different results in direct proportion to the percentage of treatment group members who did not attend a private school and control group members who did not return to public school. If only half those offered vouchers use them, and none of the control group attends a private school, then the impact, as estimated by the OLS model, will be exactly one half that of the estimated impact of actually attending a private school. As levels of non-compliance among

treatment and control group members were substantial in New York, KZ's OLS estimates are considerably lower than the 2SLS estimates we report above.

KZ provide three justifications for focusing on the effect of a voucher offer. First, they claim that the OLS estimates provide a "cleaner interpretation" of the efficacy of school vouchers. We disagree. It is not at all clear why the act of offering a voucher—as distinct from the act of using a voucher to attend a private school for one, two, or three years—should affect student achievement. Presumably, differences between treatment and control groups derive from the differential attendance patterns at public and private schools, not from the mere fact that only one group was offered vouchers. As Barnard et al. (2003b, p. 321) point out, "one could argue that the effect of attending [a private school]. . . is the more generalizable, whereas the [effect of offering will change] . . . if the next time the program is offered the compliance rates change (which seems likely!)." In short, results that isolate the impact of attending a private school provide the "cleaner interpretation" of programmatic impacts.

Second, KZ argue that the OLS model provides the better estimation of the "societal" effects of school vouchers. Presumably, the effect of an offer establishes some baseline for assessing the average gains that one can expect from a voucher intervention. This claim, however, assumes that voucher usage rates are unrelated to programmatic issues of scale, publicity, and durability. Since the New York voucher program was small, privately funded, initially limited to three years, and given only modest attention by the news media, one must make strong assumptions to infer that the voucher offer provides an accurate estimate of impacts in larger-scale programs.

Finally, KZ claim to "favor the ITT estimates" because measurement error in the endogenous variable (whether or not students attended a private school) can bias estimated 2SLS impacts "under conditions that are likely to hold." However, the direction of the bias, however, remains unclear. There is no reason to expect measurement error for the treatment group, because administrative records were used to identify students who were using the voucher to attend a private school. And for the control group, all students were assigned to public schools, unless information reported by the parent indicated otherwise. Because some of the students in the control group for whom attendance data were missing may well have been enrolled in private schools, and because 2SLS estimates increase, relative to OLS estimates, in direct proportion to the percentage of control group members who attend private schools, recovered estimates of attending a private school appear to be downwardly biased. However, this remains uncertain, inasmuch as measurement error arising from non-response is correlated with the instrument employed may introduce additional bias. Still, the issue here is quite different from the one discussed by Kane, Rouse, and Staiger (1999), who consider problems associated with systematic measurement error in self-reports of educational attainment, in which the respondents are likely to over-state their level of attainment.

We are hardly the first to emphasize 2SLS estimates within the context of a randomized field trial. For example, Alan Gerber and Don Greene (2000) employ the two-stage model when reporting results from an experiment designed to ascertain the effects of door-to-door campaigning on voter turnout in New Haven. Using the two-stage model, they observed that personal contacts with voters increases turnout by 9 percentage points, on average. Had they followed KZ's (2004) recommendation to privilege the OLS estimate, they would have found only a 3 percentage point impact, a finding that would have underestimated the actual impact of door-to-door contacts—for the

simple reason that many families assigned to treatment in their experiment were not contacted. In his class-size research, Krueger also reports without apology 2SLS estimates of attending small classes (1999).

[23] Barnard et al. (2003a, 301) exclude the kindergartner group but include the fairly small number of cases in students with missing test scores who, at baseline, were in grades 1-4. As mentioned previously, their results do not differ materially from ours.

[24] ATB reading baseline test scores were 25.4 (st. dev.=22.7) for the control group, 23.3 (st. dev.=22.5) for the treatment group. Math scores were 15.4 (st. dev.=18.2) and 15.8 (st. dev.=18.7), respectively. Nor, as a result of attrition, did the ATB group become unbalanced later on. Average composite baseline test scores were 19.3, 19.9, and 20.4 NPR points among African American students who attended the follow-up sessions in years one, two, and three, respectively; among the control group, baseline scores were 20.0, 20.4, and 21.1 NPR points for the three respective years. None of the differences are statistically significant. Although the models are more precise, point estimates barely change when baseline test scores are included in models estimating the effects of private school attendance.

[25] The difference is statistically significant at $p < .01$. These percentages refer to missing information at least one of the 16 demographic variables that KZ introduce (see below).

[26] Mismatches, however, may result from more than just administrative error. Some NATB parents may have brought to follow-up testing sessions children different from those who participated in the initial lotteries. Given that older NATBs apparently refused to take tests at baseline, they may well have resisted attending testing sessions in subsequent years. Families in the control group and

decliners in the treatment group, nonetheless, had financial incentives to attend these follow-up testing sessions—families were awarded between $50 and $100 for their continued participation in the study. Because parental surveys provided the only information available to verify the identity of these children, however, parents could have brought a child other than their own. If enough parents in the control group brought a better performing student in their child's place, this by itself could account for negative private-school impacts observed among NATBs. The problem is much less acute for ATBs, who identify themselves on both the baseline and follow-up tests.

[27] For our discussion of this issue, see Howell and Peterson (2004).

[28] KZ conclude that grade differences are minimal. As they put it, "The grade at which students are offered vouchers is unrelated to the magnitude of the treatment effect in the third year of the experiment . . . although there we find some tendency for older students to have a larger treatment effect when Kindergarten students are included." Impacts for kindergartners are negative in all three years: -0.7, -2.1, and -13.9 NPR points, respectively. By contrast, impacts for all students in the other grades, regardless of whether baseline scores are available, are significantly positive: 5.7, 4.2, and 7.5 NPR points. Interaction terms between kindergartners and treatment are significant in years one and three. Kindergartners may differ from the other cohorts or, as discussed elsewhere, the data on kindergartners may be invalid.

The possibility that voucher effects varied by grade level has been the subject of a good deal of commentary in *New York Times* coverage of our research. Reporters and columnists have conveyed the impression that impacts for African Americans varied significantly by grade level, sometimes

quoting MPR researcher David Myers to this effect (Zernike, 2000; Rothstein, 2000; Winerip, 2003).

Despite these news reports, David Myers has never identified significantly different impacts from one grade level to another in either year two or year three. Zernike quoted Myers at the end of the second year of the study as saying that positive effects were "concentrated" in a particular grade. That information, however, is not to be found in the scientific report issued at the end of the second year, which reveals no statistically significant differences in the effects by grade level (Myers, Peterson, Mayer, Chou, and Howell, 2000). In the final report issued by MPR (Mayer, Peterson, Myers, Tuttle, and Howell 2002), Myers, together with his co-authors, reported no significant differences by grade level, writing the following: "When the impact of attending private school for three years on African American student test scores was examined by grade level, we observed no statistically significant differences in the impact between grade levels (See Appendix D.) The impact for students in the younger grouping was 8.5 percentile points, and in the older grades the average impact was 9.1 points. Both impacts were statistically significant (p. 38)." In Myers and Mayer (2003) modify their position, saying "the offer of a voucher had a small positive impact on the achievement of African American students no matter which of the black definitions . . . are used; however, the impacts are concentrated among the oldest students (the grade 4 cohort)." But Myers and Mayer fail to show statistically significant differences between the grade 4 cohort and cohorts 1-3. Using MPR's revised weights, estimated private-school impacts after three years are 7.4, 3.4, 7.5, and 10.9 NPR points for African Americans in grades 1, 2, 3, and 4, respectively. None of these estimates differs significantly from the others. When grades 1–2 and 3–4 are combined, the estimates are 7.8 and 8.0 NPR points—both statistically significant at $p<.05$. In total, 127, 156, 130,

and 106 African American ATBs are included in the year three test score models for grades 1, 2, 3, and 4, respectively.

The results do not change when the African American NATBs are added to the analysis. The year three impacts from hybrid models are 6.0, 4.8, 4.0, and 11.8 NPR points for grades 1, 2, 3, and 4, respectively. None of these impacts is significantly different from the others. Once again, impacts in grades 1–2 and grades 3–4 combined are 7.9 and 7.1 NPR points—both significant at $p< 0.05$. In total, 139, 177, 139, and 122 African American ATBs and NATBs are included in the year-three test score models for each of the four respective grades.

Disagreeing with Myers and Mayer, KZ conclude that grade differences are minimal. As they put it, "The grade at which students are offered vouchers is unrelated to the magnitude of the treatment effect in the third year of the experiment . . . although there we find some tendency for older students to have a larger treatment effect when Kindergarten students are included." Indeed, as discussed above, impacts for kindergartners are negative in all three years: -0.7, -2.1, and -13.9 NPR points, respectively. By contrast, impacts for all students in the other grades, regardless of whether baseline scores are available, are significantly positive: 5.7, 4.2, and 7.5 NPR points. Interaction terms between kindergartners and treatment in test-score models come up significant in years one and three. These differences raise the question as to whether the kindergartners are genuinely different from the other cohorts or whether the data on kindergartners are invalid (see discussion in text for ways in which bias may have been introduced).

[29] Controlling for baseline test scores will not bias the estimated treatment effects as long as they are unrelated to students' assignments to treatment and control groups. As previously indicated, after

years one, two, and three, the balance of baseline test scores between the treatment and control groups appears intact (see note above).

[30] Because bootstrapped standard errors can vary from iteration to iteration, estimates presented in the tables of this paper may differ slightly.

[31] Hill, Rubin and Thomas (2002) stated that such inclusion would be important in any outcome analysis: "The high correlation commonly seen between pre-and posttest scores makes this variable a prime candidate for covariance adjustments within a linear model to take care of the remaining differences between groups" (171).

[32] Note that point estimates barely change when baseline test scores are included in models estimating the effects of private school attendance (Table 2). In addition to controlling for baseline test scores when possible, hybrid models include missing data indicators, private school status, and lottery indicators.

[33] Inasmuch as demographic information from a parent survey is more reliable than such information collected from young children, the parent, not the student, was the source of this information. In a small number of cases, a grandparent or someone else other than the parent completes the parent questionnaire. Information on the father's ethnic background was collected only at baseline.

[34] Although one parent inadvertently marked the "other" category, then wrote in "African American," no outcome test scores were available for the children.

[35] "Students . . . were added . . . [to the African American category] because a written response for the mother's race/ethnicity indicated that her race was Black, usually by writing Black/Hispanic or

Black combined with a specific Latin country" (KZ 2003, p. 27, note 25).

[36] While the key finding under discussion is whether vouchers impact the performance of African American students as distinct from others, KZ (2004) do not consistently employ a mutually exclusive classification scheme. They say: "[our analysis] treats race and Hispanic origin as mutually exclusive unless such a response was written in" (p. 27). KZ reclassified as African American parents who were identified as "black, Indian," "white/black," "African," "African Nigeria," and "black/Greek." KZ (2004) describe their procedures as follows: "Students . . . were added . . . [to this category] because a written response for the mother's race/ethnicity indicated that her race was Black, usually by writing Black/Hispanic or Black combined with a specific Latin country" (KZ 2004, p. 27, note 25).

In KZ (2003), report results for only the group they label either parent "black, non Hispanic," though they included within this category students where both parents were identified by the respondent as Hispanic. KZ (2004) may not be employing a mutually exclusive classification scheme because they define some students as African American and then, again, as Hispanic, in their analysis of vouchers on Hispanic student performance, for they say "[our analysis] treats race and Hispanic origin as mutually exclusive unless such a response was written in" (p. 27). But, of course, the key finding under discussion is whether vouchers have an impact on the performance of African American students as distinct from others.

When looking at the public schools from which Hispanic and African American students came, meanwhile, KZ (2004) treat African Americans and Hispanics as mutually exclusive. Substantively, the problem with this particular analysis is that schools are not necessarily poor or excellent, in fixed

or absolute terms, but may be appropriate or inappropriate for specific students. KZ also look at the public schools from which Hispanic and African American students came, reporting that impacts on African Americans and Hispanics coming from the same public schools differ markedly from one another. Using a system of weights for which insufficient information is available for replication to be possible, they report voucher impacts for African Americans that are a statistically significant 5 NPR points; for Hispanics an insignificant -3 NPR points.

The reported results are quite consistent with our original findings about the differential impacts of the voucher intervention on African Americans and Hispanics. Krueger and Zhu, however, use what might seem as confirmation as evidence to the contrary. More exactly, they insist that we are wrong to argue that gains observed for African Americans are due to inequities within the public sector. As they put it, "Differential characteristics of the initial public school that students with different racial backgrounds attended do *not* account for any gain in test scores that Black students may have reaped from attending private school."

In making this suggestion, KZ assume that public schools have uniform impacts on all students within them. A bad school is equally bad for African Americans and Hispanics. But schools can be good for one student without being good for another. Indeed, that is one of the central objectives of school voucher initiatives: by expanding educational options, families are able to search for schools that address the particular needs and interests of their individual child.

Schools are not necessarily poor or excellent, in fixed or absolute terms. Indeed, if teachers at schools with overlapping populations treat non-Hispanic African Americans differently from others; if they communicate differently with the mothers of African American students with the mothers of

other students; if the expectations for those from African American households are different from the expectations from other households, then the quality of public schools is not accurately ascertained when estimating test score impacts for students from different ethnic backgrounds attending overlapping schools.

[37] Edmonston, Goldstein, and Lott, eds. 1996, Appendix B: Office of Management and Budget: Statistical Directive No. 15. The Directive also calls for the listing of two other categories: "American Indian or Alaskan Native" and "Asian or Pacific Islander." The U.S. Census does not always use the combined format. When reporting results only by race, the Census includes all those who say their "race" is "black," regardless of their nationality, Hispanic or otherwise. But when reporting results within a combined table, it classifies as "Hispanic" all those who identify themselves as such, regardless of their response to a separate question on "race." Whites and blacks are then identified as white, non-Hispanic and black, non-Hispanic.

Edmonston, Goldstein, and Lott, eds. 1996, Appendix B: Office of Management and Budget: Statistical Directive No. 15. The Directive also calls for the listing of two other categories: "American Indian or Alaskan Native" and "Asian or Pacific Islander." KZ (2004) admonish Mathematica Policy Research for not using a data collection procedure recommended in this Directive, despite the fact that MPR's classification scheme is consistent with one of the options it provides. The admonishment is especially surprising, given that KZ themselves have chosen a classification scheme that fails to conform with those recommended in this very Directive.

[38] National Press Club, Washington, D.C., April 1, 2003.

[39] Myrdal (1964) explains why the African American experience, rooted in a history of slavery and

intense segregation, is unique in American society. Ethnic classifications based strictly on physical appearances ignore African Americans' distinctive history, culture, and social networks. In *The Education Gap*, for instance, we show that Hispanics, like other immigrant groups, appear to have more educational choice and suffer less from certain kinds of discrimination than African Americans.

[40] KZ 2003, p. 317, Table 2. KZ (2004) point out that the Chinese classify people by their fathers' ethnicity. Although that procedure seems inappropriate for the population under consideration in this study, it is at least a consistent coding principle, while KZ's is not. However, KZ may have solved the inconsistency problem by deciding not to create mutually exclusive categories, instead counting the same students as both African American and Hispanic. They say, on p. 31, that "these samples are not mutually exclusive" though elsewhere they claim that their analysis "treats race and Hispanic origin as mutually exclusive" (p. 27), except for the situation discussed in note 29 above.

Elsewhere, KZ (2004) clearly treat African Americans and Hispanics as mutually exclusively categories. They look at the public schools from which Hispanic and African American students came, reporting that impacts on African Americans and Hispanics coming from the same public schools differ markedly from one another. Although they conclude from this that African Americans do not attend inferior public schools, schools are not necessarily poor or excellent, in fixed or absolute terms, but may be appropriate or inappropriate for specific students.

[41] Barnard et al. (2003a), p. 305.

[42] See, for example, Phillips, Brooks-Gunn, Duncan, Klebanov, and Crane, 1998.

[43] Results are similar when ATB and NATB students are considered together.

[44] Because fathers were often not present in the household, their demographic information was missing in many cases, providing further reason for classifying according to mother's ethnicity. Indeed, demographic information is missing for 76 percent of the fathers, as compared to only 21 percent of the mothers. KZ themselves recognize the importance of mothers, noting three exceptional cases where "there is no indication the mother lived at home" and yet the child was classified according to the mother's background. When we estimate impacts for parental caretakers, these three cases are included in the estimates.

[45] Though the other two classifications are plausible, to avoid classification searching, we place primary weight on the original classification scheme, chosen prior to the conduct of the original analysis. Differences in results among the three plausible classification schemes are trivial.

[46] Eighty students had an African American father and a mother from a different ethnic background; 78 students had an African American mother and a father from a different ethnic background.

[47] All results are significant using two-tail test, except for year two results for those without baseline scores, which are significant using one-tail test.

[48] There are no missing cases for the four grade level indicators.

[49] In all, 32 percent of observations had at least one missing value on the additional covariates KZ introduce to the analysis.

[50] For the ATBs, such concerns are alleviated because we know that the baseline test scores of

treatment and control groups are balanced.

[51] The last sentence of this quote is incorrect. The possibility of swaying the estimated treatment effect is not due to chance correlations between the baseline characteristic and the outcome variable, but rather between the baseline characteristic and treatment status.

[52] In addition, only baseline test scores were mentioned, *a priori*, as a necessary benchmark when estimating achievement effects. See discussion above.

[53] KZ (2002) includes nine background controls: four indicator variables for student grade level, mother's education, log of family income, mother's employment, and gender. In KZ (2004) dropped marital status while adding controls for gifted, special education, mother born US, English-speaking household, student's age, residential mobility, mother Catholic, and welfare. With the exception of grade cohorts, none of these variables were included in Krueger's (2001b) original project proposal.

[54] Notice also that notable efficiency gains are realized simply by adding baseline test scores to the models. Standard errors drop by between 0.7 and 0.8 NPR points when baseline test scores are added to the models in all four of the ways used to classify ATB students as African American (rows 1 and 2). But no more than the most trivial gains in efficiency are realized by adding additional covariates. Even when all 28 additional covariates are added to the model, reductions in standard errors are all less than 0.1 NPR points.

The primary effect of adding covariates, instead, is to depress the point estimates on private school attendance, which drop between 1.1 and 1.5 NPR points by the time all are added to the model—a revelation that substantiates KZ's point that additional covariates may artificially "sway

the estimated treatment effect," just as it reinforces concerns about specification searching.

Table 1: Summary of Estimated Test Score Impacts for African Americans, Variously Defined

		Positive, Significant Test-Score Impacts Observed in:		
		Year One	Year Two	Year Three
	I. Simple, Transparent Models			
	A. Mother Is African American			
1-3	All students for whom baseline test scores are available, controlling for baseline scores	**	*	**
4-6	All students for whom baseline test scores are available, not controlling for baseline scores (imprecise estimation)	**	*	**
7-9	All students in grades 1-4, controlling for baseline scores when possible	**	*	**
10-12	All students in grades K-4, controlling for baseline scores when possible	**	†	*
13-15	All students in grades 1-4, not controlling for baseline scores when possible (imprecise estimation)	**	*	**
16-18	All students in grades K-4, not controlling for baseline scores when possible (imprecise estimation)	**	†	†
	B. Both Mother and Father Are African American			
19-21	All students for whom baseline test scores are available, controlling for baseline scores	**	*	**
22-24	All students for whom baseline test scores are available, not controlling for baseline scores (imprecise estimation)	**	*	**
25-27	All students in grades 1-4, controlling for baseline scores when possible	**	**	**
28-30	All students in grades K-4, controlling for baseline scores when possible	**	*	*
31-33	All students in grades 1-4, not controlling for baseline scores when possible (imprecise estimation)	**	**	**
34-36	All students in grades K-4, not controlling for baseline scores when possible (imprecise estimation)	**	†	†
	C. Parental Caretaker Is African American			
37-39	All students for whom baseline test scores are available, controlling for baseline scores	**	*	**
40-42	All students for whom baseline test scores are available, not controlling for baseline scores (imprecise estimation)	**	*	**
43-45	All students in grades 1-4, controlling for baseline scores when possible	**	*	**
46-48	All students in grades K-4, controlling for baseline scores when possible	**	†	*
49-51	All students in grades 1-4, not controlling for baseline scores when possible (imprecise estimation)	**	*	**
52-54	All students in grades K-4, not controlling for baseline scores when possible (imprecise estimation)	**	†	†
	D. Either Mother or Father Is African American (inconsistent classification scheme)			
55-57	All students for whom baseline test scores are available, controlling for baseline scores	**	†	**
58-60	All students for whom baseline test scores are available, not controlling for baseline scores (imprecise estimation)	*	--	**
61-63	All students in grades 1-4, controlling for baseline scores when possible	**	†	**
64-66	All students in grades K-4, controlling for baseline scores when possible	**	†	*
67-69	All students in grades 1-4, not controlling for baseline scores when possible (imprecise estimation)	*	--	*
70-72	All students in grades K-4, not controlling for baseline scores when possible (imprecise estimation)	†	--	--
	II. Models that Include 12 Additional Covariates: Results from the Initial KZ Specification			
	A. Mother Is African American			
73-75	All students for whom baseline test scores are available, controlling for baseline scores	**	*	**
76-78	All students in grades 1-4, controlling for baseline scores when possible	**	*	**
79-81	All students in grades K-4, controlling for baseline scores when possible	**	†	†

Table 1 Continued

		Positive, Significant Test-Score Impacts Observed in:		
		Year One	Year Two	Year Three
	B. Both Mother and Father Are African American			
82-84	All students for whom baseline test scores are available, controlling for baseline scores	**	*	**
85-87	All students in grades 1-4, controlling for baseline scores when possible	**	**	**
88-90	All students in grades K-4, controlling for baseline scores when possible	**	*	†
	C. Parental Caretaker Is African American			
91-93	All students for whom baseline test scores are available, controlling for baseline scores	**	**	**
94-96	All students in grades 1-4, controlling for baseline scores when possible	**	*	**
97-99	All students in grades K-4, controlling for baseline scores when possible	**	†	†
	D. Either Mother or Father Is African American (inconsistent classification scheme)			
100-102	All students for whom baseline test scores are available, controlling for baseline scores	**	*	**
103-105	All students in grades 1-4, controlling for baseline scores when possible	**	†	**
106-108	All students in grades K-4, controlling for baseline scores when possible	**	†	†
	III. Models that Include 28 Additional Covariates: Results from the Second KZ Specification			
	A. Mother Is African American			
109-111	All students for whom baseline test scores are available, controlling for baseline scores	**	*	**
112-114	All students in grades 1-4, controlling for baseline scores when possible	**	†	**
115-117	All students in grades K-4, controlling for baseline scores when possible	**	†	†
	B. Both Mother and Father Are African American			
118-120	All students for whom baseline test scores are available, controlling for baseline scores	**	*	**
121-123	All students in grades 1-4, controlling for baseline scores when possible	**	**	**
124-126	All students in grades K-4, controlling for baseline scores when possible	**	†	†
	C. Parental Caretaker Is African American			
127-129	All students for whom baseline test scores are available, controlling for baseline scores	**	*	**
130-132	All students in grades 1-4, controlling for baseline scores when possible	**	†	**
133-135	All students in grades K-4, controlling for baseline scores when possible	**	†	†
	D. Either Mother or Father Is African American (inconsistent classification scheme)			
136-138	All students for whom baseline test scores are available, controlling for baseline scores	**	†	**
139-141	All students in grades 1-4, controlling for baseline scores when possible	**	--	*
142-144	All students in grades K-4, controlling for baseline scores when possible	**	--	--

** significant at p<.05, two-tailed test; * significant at p<.10, two-tailed test; † at p<.10, one-tailed test. Significance tests based upon bootstrap standard errors (10,000 reps completed) that are robust to intra-family correlations.

BEST COPY AVAILABLE

Table 2: Private-School Impacts on African American Test Scores, Alternative Estimates and Efficiency Losses Resulting from Exclusion of Baseline Test Scores for Various Groups of Students

	Year One			Year Two			Year Three		
	Impact	SE	N	Impact	SE	N	Impact	SE	N
Baseline Scores in Model									
1. Students with baseline scores (ATBs, grades 1–4)	6.13**	(1.74)	622	4.16*	(2.22)	497	8.43**	(2.86)	519
No Baseline Scores in Model									
2. Students with baseline scores (ATBs, grades 1–4)	5.67**	(2.32)	622	4.36*	(2.41)	497	8.40**	(3.32)	519
3. Students with and without baseline scores (ATBs & NATBs, 1–4)	5.65**	(2.32)	695	4.24*	(2.33)	562	7.49**	(3.21)	577
4. Students with and without baseline scores (ATBs & NATBs, K–4)	4.61**	(2.07)	882	3.24†	(2.24)	722	4.88†	(3.02)	734
Hybrid Model: Baseline Scores when Possible									
5. Students with and without baseline scores (ATBs & NATBs, grades 1–4)	6.28**	(1.86)	695	3.94*	(2.21)	562	7.75**	(2.81)	577
6. Students with and without baseline scores (ATBs & NATBs, K–4)	5.15**	(1.71)	882	3.21†	(2.07)	722	5.31**	(2.70)	734

Impacts of private school attendance on test scores reported. Weighted, two-stage least squares regressions estimated, treatment status used as instrument. ** significant at p<.05, two-tailed test; * significant at p<.10, two-tailed test; † at p<.10, one-tailed test. Bootstrap standard errors (10,000 reps completed) that are robust to intra-family correlations are reported in parentheses. ATBs consist of students for whom baseline test scores are available; NATBs consist of students for whom no baseline test scores are available. First set of models include as covariates private school status, baseline test scores, and lottery indicators; the second set include only private school status and lottery indicators; the hybrid model includes private school status, baseline test scores (interacted with a dummy variable for students with baseline test scores), the dummy variable for students with baseline test scores, and lottery indicators.

BEST COPY AVAILABLE

Table 3: Test Score Impacts for African Americans, Variously Defined

	MOTHER AFRICAN AMERICAN			MOTHER & FATHER AFRICAN AMERICAN			PARENTAL CARETAKER AFRICAN AMERICAN[1]			MOTHER OR FATHER AFRICAN AMERICAN		
	Impact	SE	N	Impact	SE	N	Impact	SE	N	Impact	SE	N
Students with Baseline Scores (ATBs)												
Year One	6.13**	(1.72)	622	5.78**	(1.84)	587	6.18**	(1.74)	624	5.29**	(1.75)	667
Year Two	4.16*	(2.23)	497	4.13*	(2.29)	469	4.17*	(2.24)	498	3.28†	(2.18)	533
Year Three	8.43**	(2.86)	519	8.05**	(2.93)	485	8.36**	(2.87)	520	7.64**	(2.83)	554
Students with and without Baseline Scores (ATBs & NATBs)												
Year One	5.15**	(1.76)	882	5.00**	(1.81)	839	5.20**	(1.74)	884	4.00**	(1.72)	946
Year Two	3.21†	(2.08)	722	3.56†	(2.14)	683	3.24†	(2.07)	723	2.66†	(2.03)	771
Year Three	5.31**	(2.71)	734	5.08*	(2.82)	687	5.27**	(2.70)	735	4.45*	(2.65)	785

Impacts of private school attendance on test scores reported. Weighted, two-stage least squares regressions estimated; treatment status used as instrument. ** significant at p<.05, two-tailed test; * significant at p<.10, two-tailed test; † at p<.10. Bootstrap standard errors (10,000 reps completed) that are robust to intra-family correlations are reported in parentheses. Models for students with baseline test scores control for baseline scores and lottery indicators; models for all students control for test scores when possible, an indicator variable for missing baseline scores, and lottery indicators. Mother's ethnicity determined on the basis of baseline, year two, and year three surveys; father's ethnicity determined on the basis of baseline surveys only. When accounting for the ethnicity of both parents, if missing, mother [father] assumed African American when father [mother] African American. ATBs consist of students for whom baseline test scores are available; NATBs consist of students for whom no baseline test scores are available.

[1] Mother assumed to be the primary caretaker of the child's education except in those cases where the child lives only with the father.

Table 4: Year Three Test Score Impacts for African Americans, Variously Defined, With and Without Baseline Test Scores (Estimates Obtained from Simple/Transparent Models and from Specification Searches)

	Student with Baseline Test Scores (ATBs)				Students with and without Baseline Test Scores (ATBs & NATBs)			
	Mother African American	Both Mother & Father AA	Parental Caretaker African American[1]	Either Mother or Father AA	Mother African American	Both Mother & Father AA	Parental Caretaker African American[1]	Either Mother or Father AA
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Transparent Model (no baseline test scores)	8.40** (3.35)	7.91** (3.44)	8.41** (3.31)	7.10** (3.24)	4.88' (3.00)	4.69' (3.16)	4.90' (3.04)	3.53 (2.96)
Transparent Model (with baseline test scores)[2]	8.42** (2.82)	8.05** (2.91)	8.36** (2.84)	7.64** (2.80)	5.31** (2.70)	5.08* (2.79)	5.27* (2.72)	4.45* (2.64)
1st Search, Controls for:								
four grade levels[3]	7.88** (2.80)	7.18** (2.89)	7.84** (2.83)	7.43** (2.76)	4.79* (2.74)	4.31' (2.77)	4.79* (2.70)	4.23' (2.61)
Plus mother's education	7.80** (2.82)	7.15** (2.86)	7.77** (2.79)	7.32** (2.73)	4.82* (2.69)	4.39' (2.74)	4.82* (2.68)	4.19' (2.59)
Plus log income	7.79** (2.80)	7.17** (2.85)	7.76** (2.83)	7.35** (2.69)	4.82* (2.75)	4.38' (2.77)	4.82* (2.73)	4.18' (2.63)
Plus student's gender	7.74** (2.85)	7.16** (2.89)	7.71** (2.83)	7.36** (2.73)	4.81* (2.78)	4.46' (2.84)	4.80* (2.79)	4.15' (2.64)
Plus employment status	7.85** (2.85)	7.23** (2.89)	7.81** (2.82)	7.79** (2.71)	4.40' (2.86)	4.44' (2.80)	4.40' (2.83)	3.88' (2.65)
2nd Search, Adds Controls:								
welfare	7.95** (2.88)	7.36** (2.94)	7.91** (2.90)	7.87** (2.81)	4.40' (2.83)	4.52' (2.87)	4.39' (2.86)	3.84' (2.76)
Plus mother born US	7.80** (2.82)	7.11** (2.83)	7.76** (2.85)	7.74** (2.72)	4.20' (2.81)	4.26' (2.82)	4.19' (2.78)	3.51' (2.68)
Plus residential mobility	7.98** (2.79)	7.07** (2.77)	7.94** (2.82)	7.70** (2.68)	4.32' (2.82)	4.15' (2.78)	4.31' (2.79)	3.55' (2.66)
Plus English spoken home	7.50** (2.77)	6.61** (2.75)	7.46** (2.79)	7.23** (2.67)	3.96' (2.74)	3.81' (2.74)	3.95' (2.73)	3.19 (2.64)
Plus mother Catholic	7.23** (2.79)	6.40** (2.75)	7.19** (2.76)	7.04** (2.70)	3.82' (2.73)	3.80' (2.73)	3.81' (2.70)	3.11 (2.63)
Plus student's age	7.19** (2.77)	6.37** (2.74)	7.15** (2.76)	7.03** (2.69)	3.83' (2.75)	3.79' (2.74)	3.81' (2.75)	3.11 (2.64)
Plus student gifted	7.10** (2.85)	6.28** (2.82)	7.06** (2.80)	6.96** (2.71)	3.62' (2.71)	3.61' (2.75)	3.60' (2.75)	3.06 (2.62)
Plus student special ed.	7.20** (2.77)	6.39** (2.83)	7.16** (2.76)	6.90** (2.66)	3.55' (2.72)	3.66' (2.74)	3.53' (2.74)	2.91 (2.60)
(N)	519	485	520	554	734	687	735	785

Impacts of private school attendance on test scores. Weighted, two-stage least squares regressions estimated; treatment status used as instrument. ** significant at p<.05, two-tailed test; * significant at p<.10, two-tailed test; † at p<.10. Bootstrap standard errors (10,000 reps completed) that are robust to intra-family correlations are reported in parentheses. Mother's ethnicity determined on the basis of baseline, year two, and year three surveys; father's ethnicity determined on the basis of baseline surveys only. When accounting for the ethnicity of both parents, if missing, mother [father] assumed African American when father [mother] African American. All models include as covariates private school status and revised lottery indicators. Covariates then added cumulatively, so that final row includes test scores and all 16 additional demographic controls and all 12 missing value indicators that are used in the Krueger/Zhu estimations. Among African American mothers, 6.5 percent of cases are missing for mother's education, 7.6 percent for income, 3.3 for gender, 2.6 for employment status, 10.9 for welfare, 2.1 for born U.S., 2.9 for residential mobility, 3.0 for English spoken at home, 7.7 for Catholic, 4.4 for age, 3.5 for gifted, and 3.9 for special education. The first five rows of additional controls are those covariates included in KZ (2002). The last 8 rows are those covariates included in KZ (2004). KZ (2002) also included marital status, but then was dropped from later analyses.)

[1] Mother assumed to be the primary caretaker of the child's education except in those cases where the child lives only with the father.

[2] Models for students with baseline test scores include as covariates private school status, baseline scores and lottery indicators; models for all students include as covariates private school status, baseline test scores when possible, an indicator variable for missing baseline scores, and lottery indicators.

[3] Three grade-level indicator variables are included for models that only include ATBs.

BEST COPY AVAILABLE

REFERENCES

Achen, Christopher. 1986. *The Statistical Analysis of Quasi-Experiments*. Berkeley: University of California Press.

Barnard, John, Constantine E. Frangakis, Jennifer L. Hill, and Donald B. Rubin. 2003a. "Principal Stratification Approach to Broken Randomized Experiments: A Case Study of School Choice Vouchers in New York City." *Journal of the American Statistical Association.* 98 (June): 299-311.

Barnard, John, Constantine E. Frangakis, Jennifer L. Hill, and Donald B. Rubin. 2003b. "Rejoinder." *Journal of the American Statistical Association.* 98 (June): 320-23.

Chubb, John E. and Terry M. Moe. 1990. *Politics, Markets, and America's Schools* Brookings.

Coleman, James S., Thomas Hoffer, and Sally Kilgore. *High School Achievement.* New York: Basic Books, 1982.

Corporation for the Advancement of Policy Evaluation with Mathematic Policy Research. 1996. "Evaluation of the New York City Scholarship Program," Proposal submitted to Phoebe Cottingham, Senior Program Officer, Smith Richardson Foundation. December 11.

Corporation for the Advancement of Policy Evaluation with Mathematica Policy Research. 1997. "Evaluation of the New York City Scholarship Program, Technical and Cost Proposal." Proposal submitted to Phoebe Cottingham, Senior Program Officer, Smith Richardson Foundation. November 24.

Evans, William N. and Robert M. Schwab, 1993. "Who Benefits from Private Education? Evidence from Quantile Regressions." Department of Economics, University of Maryland .

Edmonston, Barry, Joshua Goldstein, and Juanita Tamayo Lott, eds., 1996. *Spotlight on Heterogeneity: The Federal Standards for Racial and Ethnic Classification, Summary of a Workshop.* National Academy Press.

Figlio, David N. and Joe A. Stone. 1999. "Are Private Schools Really Better?" *Research in Labor Economics.* JAI Press, Inc. 1 (18): 115-140.

Jencks,Christopher.1985. "How Much Do High School Students Learn?" *Sociology of Education*. 58:128-35.

Gerber, Alan S. and Donald P. Green. 2000. "The Effects of Canvassing, Telephone Calls, and Direct Mail on Voter Turnout: A Field Experiment," *American Political Science Review,* 94 September: 653-63.

Godwin R. Kenneth. 2002. "Choice Words," *Education Next*, Fall, 2002, 2(3).

Greene, Jay P., Paul E. Peterson, and Jiangtao Du. 1998. "School Choice in Milwaukee: A Randomized Experiment." In *Learning from School Choice, eds.* Paul E. Peterson and Bryan C. Hassel. Brookings. Pp. 335-56.

Grogger, Jeffrey and Derek Neal, 2000. "Further Evidence on the Effects of Catholic Secondary Schooling." *Brookings-Wharton Papers on Urban Affairs: 2000.* (Washington, D.C.: Brookings.)

Hill, Jennifer, Donald Rubin, and Neal Thomas. 2002. "The Design of the New York School Choice Scholarship Program Evaluation," in *Donald Campbell's Legacy*, ed. L. Bickman, Newbury Park, CA: Sage Publications.

Howell, William G., Paul E. Peterson with Patrick J. Wolf and David E. Campbell. 2002. *The Education Gap: Vouchers and Urban Schools.* Brookings.

Howell, William G., Patrick Wolf, David Campbell, and Paul E. Peterson. 2002. "School Vouchers and Academic Performance: Results from Three Randomized Field Trials." *Journal of Policy Analysis and Management.* 21(2): 191-218.

Howell, William G. and Paul E. Peterson. 2004[forthcoming]. "The Use of Theory in Randomized Field Trials: Lessons from School Voucher Research on Disaggregation, Missing Data, and the Generalizability of Findings," *The American Behavioral Scientist.*

Krueger, Alan B. 1999. "Experimental Estimates of Education Production Functions." *Quarterly Journal of Economics.* 114 (May): 497-532.

Krueger, Alan B. 2001a. *Education Matters: Selected Essays.* Elgar.

Krueger, Alan B. 2001b. "Data License and Confidentiality Agreement Reanalysis of the Data Used in "School Choice in New York City after Two Years: An Evaluation of the School Choice Scholarships Program." Project proposal submitted to Joanne Pfleiderer, Director of Communications, Mathematica Policy Research. May 9.

Krueger, Alan B. and Pei Zhu. 2002. "Another Look at the New York City School Voucher Experiment." A paper prepared for the Conference on Randomized Experimentation in the Social Sciences, Yale University. (August 16).

Krueger, Alan B. and Pei Zhu. 2003. "Comment [on Barnard, Frangakis, Hill and Rubin, 2003]," *Journal of the American Statistical Association.* 98(June): 314-318.

Krueger, Alan and Pei Zhu. 2004[forthcoming]. "Another Look at the New York City School Voucher Experiment." *American Behavioral Scientist.*

Mayer, Daniel P., Paul E. Peterson, David E. Myers, Christina Clark Tuttle, and William G. Howell. 2002. "School Choice in New York City After Three Years: An Evaluation of the School Choice Scholarships Program, Final Report." Research Paper 8404-045. Princeton, New Jersey: Mathematica Policy Research, February.

Myers, David E. and Daniel P. Mayer. 2003. "Comments on a 'Another Look at the New York City Voucher Experiment.'" Memorandum. Washington, D. C. Mathematica Policy Research. April 1.

Myers, David, Paul E. Peterson, Daniel Mayer, Julia Chou, and William G. Howell. 2002. "School Choice in New York City after Two Years: An Evaluation of the School Choice Scholarships Program," Program on Education Policy and Governance, Harvard University, Report 00-17.

Myrdal, Gunner. 1964. *An American Dilemma.* (McGraw Hill).

Neal, Derek. 2003. "Investment Planning." *Education Next.* 3 (Winter): 85.

Peterson, Paul E., William G. Howell, Patrick J. Wolf, and David E. Campbell. 2003. "School Vouchers: Results from Randomized Experiments." In *The Economics of School Choice,* ed. Caroline M. Hoxby. Chicago: University of Chicago.

Phillips, Meredith, Jeanne Brooks-Gunn, Greg J. Duncan, Pamela Klebanov, Jonathan Crane. 1998. "Family Background, Parenting Practices, and the Black-White Test Score Gap," In, eds., *The Black-White Test Score Gap,* eds. Christopher Jencks and Meredith Phillips. Washington, D. C.: Brookings. Pp.103-48.

Rothstein, Richard. "Judging Voucher Merits Proves a Difficult Task," *New York Times*, December 13, 2000

Rouse, Cecilia Elena. 2000. "School Reform in the 21st Century: A Look at the Effect of Class Size and School Vouchers on the Academic Achievement of Minority Students." Princeton University Working Paper 440.

Sanders, William. 2002. *Catholic Schools: Private and Social Effects.* Kluwer.

Winerip, Michael. 2003. "What Some Much-Noted Data Really Showed about Vouchers," *New York Times*, May 7.

Zellner, A. 1984. *Basic Issues in Econometrics*. Chicago: University of Chicago Press

Zernike, Kate 2000. "New Doubt Is Cast on Study that Backs Voucher Effects," *New York Times*, September 15.

Appendix: Reply to the Krueger and Zhu Rejoinder

For the 2003 APPAM meetings, Krueger and Zhu (KZ) made available a rejoinder to this paper. Unfortunately, it offers little new information on the points under discussion, and it makes strong charges without adequate documentation. The authors pretend to be unaware of information in the public domain and available to them (indeed, sent to them), but quote extensively from private emails and preliminary drafts of papers. In this appendix, we summarize and respond to various KZ assertions made in this rejoinder:

KZ Assertion No. 1: *We could not replicate Peterson and Howell's findings. "They did not cooperate in an attempt to understand the differences," and hence "we recommend skepticism concerning their analysis."*

Our Response: KZ's claim is misleading. In the text, KZ imply that they cannot replicate our findings in general; but in a qualifying note, they identify discrepancies only to the complex equations contained in Table 4 of our paper. Their re-analyses of these models (the only ones they have released) nonetheless confirm our four basic arguments: (1) impacts for African American students for whom baseline test score information is available (Table A1, columns 1-4) are always significantly positive, whether one defines a child's ethnicity by the mother or other ways; (2) efficiency gains and losses associated with adding covariates beyond baseline test scores are minor and inconsistent; (3) the additional covariates consistently depress the point estimates; and (4) the estimates from hybrid models for African American students that control just for baseline test scores are significantly positive, regardless of one's definition of African American (Table A1 columns 5-8, row 2). With information from KZ's reanalysis now available, it is more clear than ever that unless one is willing to ignore baseline test-score

information (even when available), or one loads the regression with covariates that do not materially improve the precision of the estimates, all models reveal statistically significant, positive impacts on the year three test scores of African Americans. Compare the following results from models that control for baseline test scores whenever possible:

	Mother AA	**Both Mother & Father AA**	**Parental Caretaker AA**	**Either Mother or Father AA**
Students for whom baseline test scores are available				
Our estimates	8.4**	8.1**	8.4**	7.6**
KZ's estimates	8.4**	7.4**	7.6**	7.6**
All students, regardless of whether baseline test scores are available				
Our estimates	5.3**	5.1*	5.3*	4.5*
KZ's estimates	5.4**	4.0*	4.8*	4.2*

**** p<.05, two-tailed test; * p<.10. Two-stage least squares models estimated; treatment status used as instrument. All models control for baseline test scores, whenever possible, as well as lottery indicators.**

It is the case that we differ somewhat in the number of additional covariates that one must add before the estimates cross standard thresholds of statistical significance. But we put little weight on these results and, judging from the fact that they do not appear in Krueger and Zhu (2003), neither do they.

Still, clear errors are apparent in KZ's replication efforts. For instance, their sample sizes for "parental caretaker African American" are 15 percent smaller than those for "mother African American," despite the fact that our definition of the former population requires a sample size that is equal to or larger than that for the latter population. We repeatedly invited KZ to send us their programming codes so that such errors could be avoided; unfortunately, they did not do so.

Separately, we have not been able to replicate many of KZ's findings, but because differences uncovered were small, it seemed distracting to raise this issue in a public forum.

KZ assertion No. 2: *Our sample sizes differ from those "of an unpublished table prepared to accompany" a presentation by David Myers and Daniel Mayer (two analysts at Mathematica Policy Research).*

Our response: Our sample sizes exactly match those contained in the document that Myers and Mayer made publicly available after their presentation at the National Press Club in April 2003 (Myers and Mayer 2003). As best we can tell, KZ's assertion seems to be based on Myers and Mayer's preliminary work.

KZ assertion No. 3: When impacts are estimated without baseline test scores for all African Americans, "The effect of vouchers on composite scores is less than a tenth of a standard deviation and statistically insignificant at the 10% level in Years 2 and 3."

Our response: The effect is statistically significant at the 5 percent level once baseline test scores are controlled, whenever possible, even when all students participating in the evaluation are included in the analysis. And KZ are speaking of the effect of "the offer of a voucher," not the "effect of vouchers," an error that pervades their rejoinder. In fact, the effect of private school attendance on the test scores of African Americans varies between a quarter and two-fifths of a standard deviation, depending on the estimation. When precisely estimated, effects are always statistically significant – even when students without baseline test scores are included in the analysis – provided that one controls for baseline test scores whenever possible.

KZ assertion No. 4: *An employee of the Office of Management and Budget (OMB) states in a private email to KZ that questionnaires should permit the reporting of multiple races.*

Our response: The quotation from the private email is hardly relevant, because it discusses questionnaire design, not the organization of data in analyses that compare and contrast African Americans, Hispanics, and other ethnic groups. The email, further, ignores the fact that respondents who wished to provide two ethnicities could check the "other" box and identify themselves however they liked. Finally, the privately expressed opinion of an employee does not constitute OMB policy.

As to questionnaire design, OMB policy, as stated in Statistical Directive No. 15 (the directive that KZ rely upon extensively in their original paper), is quite clear. One can either use separate questions or a combined format. If one uses a combined format (the approach we followed) one should have at least the following categories: black (non-Hispanic), white (non-Hispanic), and black. Nothing in the directive prohibits the use of additional categories, nor does OMB pretend to dictate the way in which scientists should analyze their data. On the contrary, Statistical Directive No. 15 explicitly says, in its opening paragraphs, that its guidelines are intended to inform the work of government agencies, not scientific investigations.

Nonetheless, government agencies typically organize data in a manner consistent with our practice, and contrary to that recommended by KZ. For example, the recently released longitudinal survey, Early Childhood Longitudinal Study (ECLS) administered by the Department of Education, organized data for secondary analyses into the same categories that we did: black (non-Hispanic), white (non-Hispanic), Hispanic, and, of course, other groups. Plainly, the Department of Education's decision to release data in this form does not violate OMB Statistical Directive No. 15.

Drawing upon ECLS data, Fryer and Levitt (Roland G. Fryer and Steven D. Levitt, "Understanding the Black-White Test Score Gap in the First Two Years of School." *Review of Economics and Statistics,* forthcoming) report that African American students lose ground relative to whites as they move through the first two years of schooling. Meanwhile, Hispanics gain ground relative to whites. These contrasting results for African American and Hispanic students are based on the same classification system employed in our research. These authors thank Alan Krueger for his assistance with this paper, though nothing in the paper indicates that Krueger objected to the classification system the authors employed.

KZ assertion No. 5: *"If someone marked other and wrote in Hispanic/Black, we appropriately classified that person in both categories, consistent with the OMB guidelines."*

Our response: OMB guidelines make no recommendations as to how data are to be analyzed by social scientists, who generally employ mutually exclusive categories, especially when analyzing similarities and differences between groups. For further discussion see discussion of Assertion No. 4, above.

KZ Assertion No. 6: *An employee at Mathematica Policy Research states in a private email to Krueger that data collection protocols for kindergartners make it unlikely that children were misidentified.*

Our Response: That an employee of a for-profit, commercial firm defends its procedures does not make a problem disappear.

KZ assertion No. 7: *Peterson and Howell did not from the beginning plan to control for baseline test scores. The project proposal they submitted to foundations never indicated that they were interested in doing so. They only controlled for them later on to get results they desired.*

Our response: A project proposal prepared five months before outcome data had been collected said the following: "The math and reading achievement tests completed by students [at baseline] will provide a benchmark against which to compare students' future test scores" (Corporation for the Advancement of Policy Evaluation with Mathematica Policy Research, Inc., 1997.) Since the project proposal was prepared for a lay board, the language used was non-technical, but the intention is clear.

Another paper prepared at an early stage of the research and later published as Hill, Rubin and Thomas (2002) stated that inclusion of baseline test scores in outcome analyses was critically important: "The high correlation commonly seen between pre-and posttest scores makes this variable a prime candidate for covariance adjustments within a linear model to take care of the remaining differences between groups" (171). This paper was written before Hill and her colleagues had conducted any outcome analyses. Their later work (Barnard et al 2003) also finds significantly positive impacts on the test scores of African American students, confirming our original results for the first outcome year.

KZ assertion No. 8: "Because they found a negative effect of vouchers in Washington, D.C. after three years (see Table 6-1 of HP, 2002), we see no compelling reason to conduct a one-tailed test."

Our Response: KZ fail to report that the year three effect in Washington, D.C. was statistically insignificant, a startling fact given that the issue under consideration is statistical significance.

In New York, all results using an appropriate methodology survive a two-tailed test. Nonetheless, a one-tail test is appropriate when prior research (or theory) suggests that effects will either be null or point in only one direction. As note 8 to this paper states: "Given the sheer number of studies of private schools that find positive achievement and attainment effects for African American students (Chubb and Moe 1990; Coleman, Hofer, and Kilgore 1982; Evans and Schwab 1993; Figlio and Stone 1999; Grogger and Neal 2000; Jencks 1985; Neal 1997; Rouse 2000), and given that no major study—to our knowledge—has found that private schools adversely affect the education of African American students, a one-tailed test is not inappropriate."

KZ assertion No. 9: *Peterson, in his study of voucher impacts in Milwaukee, said that baseline test scores are unnecessary in randomized field trials.*

Our response: Here is what Peterson and his co-authors actually said in Greene, Peterson and Du, 1998, the publication cited by KZ: "The conclusions that can be drawn from our study are . . . restricted by limitations of the data. . . . . The percentage of missing cases is especially large when one introduces controls for . . . pre-experimental test scores. But given the consistency and magnitude of the findings . . . they suggest the desirability of further randomized experiments capable of reaching more precise estimates of efficiency gains through privatization. Randomized experiments are underway in New York City, Dayton, and Washington, D.C. If the evaluations of these randomized experiments minimize the number of missing cases and collect

pre-experimental data for all subjects. . . , they could . . . provide more precise estimates of potential efficiency gains" (p. 351).

KZ Assertion No. 10: *Peterson and Howell never respond to our criticism of their two-stage model.*

Our response: We, in fact, did respond (Peterson and Howell 2003; Howell and Peterson 2004). Copies of these papers have been made available to KZ. Unfortunately, they have not responded substantively. Nor do KZ reply to Barnard et al. (2003b, p. 321), who reject their argument that the estimated effect of a voucher offer is more generalizable than the estimated effect of actually attending a private school. As Barnard et al point out, "one could argue that the effect of attending [a private school]. . . is the more generalizable, whereas the [effect of offering will change] . . . if the next time the program is offered the compliance rates change (which seems likely!)."

The response to which KZ have yet to respond is as follows:

> Krueger and Zhu argue that programmatic effects are best understood by examining the impact of being offered a voucher rather than the impact of actually attending a private school. The first impact, known as intent-to-treat (ITT), is estimated by ordinary least squares (OLS); the second by a two-stage model (2SLS), which uses the randomized assignment to treatment and control conditions as an instrument for private school attendance. Almost all of the estimates Krueger and Zhu provide are based on the OLS model.
>
> To ascertain the statistical significance of programmatic effects, it makes no difference which model is estimated. Both yield identical results. If, however, one is interested in the magnitude of an intervention's impact, not just its statistical significance, then the choice of models is critical. The two estimators will yield different results in direct proportion to the percentage of treatment group members who did not attend a private school and control group members who did not return to public school. If only

half those offered vouchers use them, and none of the control group attends a private school, then the impact, as estimated by the OLS model, will be exactly one half that of the estimated impact of actually attending a private school. As levels of non-compliance among treatment and control group members were substantial in New York, Krueger and Zhu's OLS estimates are considerably lower than the 2SLS estimates we report above.

Krueger and Zhu provide three justifications for focusing on the effect of a voucher offer. First, they claim that the OLS estimates provide a "cleaner interpretation" of the efficacy of school vouchers. We disagree. It is not at all clear why the act of offering a voucher—as distinct from the act of using a voucher to attend a private school for one, two, or three years—should affect student achievement. Presumably, differences between treatment and control groups derive from the differential attendance patterns at public and private schools, not from the mere fact that only one group was offered vouchers. Results that isolate the impact of attending a private school provide the "cleaner interpretation" of programmatic impacts.

Second, Krueger and Zhu argue that the OLS model provides the better estimation of the "societal" effects of school vouchers. Presumably, the effect of an offer establishes some baseline for assessing the average gains that one can expect from a voucher intervention. This claim, however, assumes that voucher usage rates are unrelated to programmatic issues of scale, publicity, and durability. Since the New York voucher program was small, privately funded, initially limited to three years, and given only modest attention by the news media, one must make strong assumptions to infer that the voucher offer provides an accurate estimate of impacts in larger-scale programs.

Finally, Krueger and Zhu claim to "favor the ITT estimates" because measurement error in the endogenous variable (whether or not students attended a private school) can bias estimated 2SLS impacts "under conditions that are likely to hold." However, the direction of the bias, however, remains unclear. There is no reason to expect measurement error for the treatment group, because administrative records were used to identify students who were using the voucher to attend a private school. And for the control group, all students were assigned to public schools, unless information reported by the parent indicated otherwise. Because some of the students in the control group for whom attendance data were missing may well have been enrolled in private schools, and because 2SLS estimates increase, relative to OLS estimates, in direct proportion to the percentage of control group members who attend private schools, recovered estimates of attending a private school appear to be downwardly biased.

However, this remains uncertain, inasmuch as measurement error arising from non-response is correlated with the instrument employed may introduce additional bias. Still, the issue here is quite different from the one discussed by Thomas Kane, Cecilia Rouse, and Douglas Staiger, who consider problems associated with systematic measurement error in self-reports of educational attainment, in which the respondents are likely to over-state their level of attainment. ("Estimating Returns to Schooling When Schooling is Misreported," Working Paper #419, Industrial Relations Section, Princeton University).

We are hardly the first to emphasize 2SLS estimates within the context of a randomized field trial. For example, Alan Gerber and Don Greene employ the two-stage model when reporting results from an experiment designed to ascertain the effects of door-to-door campaigning on voter turnout in New Haven. Alan S. Gerber and Donald P. Green, "The Effects of Canvassing, Telephone Calls, and Direct Mail on Voter Turnout: A Field Experiment," *American Political Science Review,* Vol. 94, (September 2000), pp. 653-63. Using the two-stage model, they observed that personal contacts with voters increases turnout by 9 percentage points, on average. Had they followed Krueger and Zhu's recommendation to privilege the OLS estimate, they would have found only a 3 percentage point impact, a finding that would have underestimated the actual impact of door-to-door contacts—for the simple reason that many families assigned to treatment in their experiment were not contacted. In his class-size research, Krueger also reports without apology 2SLS estimates of attending small classes (1999).

KZ assertion No. 11: *Peterson and Howell do not follow their own recommendation as to how Hispanics should be classified.*

Our response: Any errors in classification apply to a trivial number of cases. In any case, the point is irrelevant, since there is no disagreement with respect to the impact of private school attendance on Hispanic performance.

KZ Assertion No. 12: *Howell and Peterson (2002) made an error in an early draft of their paper.*

Our Response: The paper KZ refer to is marked "preliminary paper, comments invited" and was shared with a few scholars for their comments. It was not distributed generally. Nor do we understand how the reader is made to benefit by KZ identifying errors in early drafts of papers. Over the past year, we have identified numerous errors in both KZ's original paper and in their rejoinder, some of which they have corrected. Pointing out errors that KZ have subsequently corrected would only cloud the fundamental issue at stake in this exchange –

namely, whether African Americans who switched from public to private schools in New York City posted positive test score gains. The overwhelming weight of the evidence suggests that, in fact, they did

Table A1

	Student with Baseline Test Scores (ATBs)				Students with and without Baseline Test Scores (ATBs & NATBs)			
	Mother African American	Both Mother & Father AA	Parental Caretaker African American	Either Mother or Father AA	Mother African American	Both Mother & Father AA	Parental Caretaker African American1	Either Mother or Father AA
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Transparent Model (no baseline test scores)	8.40** (3.42)	7.50** (3.49)	6.66** (3.60)	7.15** (3.22)	4.94† (3.08)	3.94 (3.26)	3.83 (3.35)	3.49 (2.86)
Transparent Model (with baseline test scores)	8.43** (2.89)	7.44** (3.15)	7.62** (3.16)	7.57** (2.66)	5.38** (2.77)	4.02* (2.97)	4.80* (3.01)	4.23* (2.52)
1st Search, Controls for:								
four grade levels	7.88** (2.83)	6.77** (3.02)	7.27** (3.12)	7.45** (2.63)	4.86* (2.77)	3.53 (2.90)	4.33* (2.97)	4.05 (2.51)
Plus mother's education	7.80** (2.80)	6.64** (3.08)	7.19** (3.06)	7.40** (2.60)	4.87* (2.75)	3.51 (2.93)	4.20* (2.96)	4.05 (2.49)
Plus log income	7.77** (2.80)	6.63** (3.14)	7.19** (3.08)	7.40** (2.62)	4.87* (2.76)	3.46 (2.97)	4.20* (3.01)	4.04 (2.51)
Plus student's gender	7.74** (2.82)	6.73** (3.17)	7.21** (3.10)	7.41** (2.62)	4.91* (2.81)	3.40 (3.02)	4.22* (3.03)	4.01 (2.55)
Plus employment status	7.64** (2.79)	6.55** (3.13)	7.06** (3.09)	7.33** (2.63)	4.67† (2.82)	2.98 (3.02)	3.74 (3.04)	3.73 (2.55)
2nd Search, Adds Controls:								
welfare	7.75** (2.84)	7.03** (3.26)	7.33** (3.15)	7.48** (2.69)	4.71† (2.85)	3.22 (3.10)	3.70 (3.09)	3.73 (2.59)
Plus mother born US	7.73** (2.81)	6.91** (3.16)	7.09** (3.07)	7.21** (2.65)	4.68† (2.85)	3.02 (3.06)	3.68 (3.05)	3.49 (2.54)
Plus residential mobility	7.80** (2.77)	7.18** (3.17)	7.12** (3.05)	7.22** (2.63)	4.73† (2.75)	3.19 (3.09)	3.75 (3.04)	3.54 (2.52)
Plus English spoken home	7.60** (2.74)	6.60** (3.17)	6.56** (3.01)	6.96** (2.62)	4.56 (2.67)	2.61 (3.02)	3.21 (3.00)	3.31 (2.49)
Plus mother Catholic	7.16** (2.72)	6.47** (3.19)	6.35** (2.99)	6.69** (2.63)	4.07 (2.64)	2.65 (2.95)	2.89 (2.96)	2.90 (2.47)
Plus student's age	7.16** (2.73)	6.47** (3.20)	6.38** (2.99)	6.68** (2.63)	4.03 (2.63)	2.44 (2.94)	2.82 (2.94)	2.74 (2.48)
Plus student gifted	6.97** (2.79)	6.51** (3.27)	6.07** (3.06)	6.44** (2.67)	3.81 (2.65)	2.13 (2.97)	2.58 (2.93)	2.67 (2.47)
Plus student special ed.	7.10** (2.75)	6.77** (3.21)	6.21** (3.14)	6.44** (2.62)	3.73 (2.63)	2.20 (2.99)	2.65 (2.91)	2.48 (2.44)
(N)	519	384	444	573	733	533	626	811

BEST COPY AVAILABLE

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC

UD 035962

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Latest Results from the New York City Voucher Experiment	
Author(s): See attached	
Corporate Source: PEPG	Publication Date: 11-2003

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Level 1

☒

Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) and paper copy.

The sample sticker shown below will be affixed to all Level 2A documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

Level 2A

☐

Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only

The sample sticker shown below will be affixed to all Level 2B documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

Level 2B

☐

Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, → please

Signature: Paul E Peterson	Printed Name/Position/Title: Paul E Peterson / Director
Organization/Address: 79 JFK St. T-306 Cambridge, MA 02138	Telephone: 617-495-8512 FAX: 617-496-4428
	E-Mail Address: ppeterson@hks.harvard.edu Date: 11/25/03

(over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, or, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
4483-A Forbes Boulevard
Lanham, Maryland 20706

Telephone: 301-552-4200
Toll Free: 800-799-3742
FAX: 301-552-4700
e-mail: info@ericfac.piccard.csc.com
WWW: http://ericfacility.org